Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 48 (2015) 58 – 64

Procedia Computer Science

International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2015)

Conference Organized by Interscience Institute of Management and Technology,

Bhubaneswar, Odisha, India

Influence of lexical, syntactic and structural features and their combination on Authorship Attribution for Telugu Text

S.NagaPrasad[a], Dr.V.B.Narsimha[b], Dr.P.Vijayapal Reddy[c], Dr.A.Vinaya Babu[d]

[a.] *Research Scholar, Dept. Of CSE, Aacharya Nagarjuna University, Guntur, India.*
[b.] *Assistant Professor, Dept. Of CSE, University College of Engineering, Osmania University, Hyderabad, India.*
[c] *Professor, Dept. Of CSE, Gokaraju Rangaraju Institute of Engineering & Technology, Hyderabad, India.*
[d.] *Professor, Dept. Of CSE, J.N.T.U. College of Engineering, Hyderabad, India*

Abstract

Authorship attribution (AA) is the task of identifying author of an unknown text from the known author set. Authorship Attribution can be viewed is a problem of text classification. AA is based on the classification of documents on author writing style rather than the topic of the text. In this paper experimental evaluations were carried out on Telugu text for Authorship Attribution using various types of features and their combinations. Feature vectors were formed for the training set using lexical, syntactic and structural features and their combinations. Learned model was generated for each these vectors and performance of the learned model is calculated using F1 metric and accuracy. More number of features can slow down the model performance. Features which are not relevant or not more relevant were eliminated from the feature vectors using chi-square metric. Support Vector Machine (SVM) algorithm is used as a classifier to generate the learned model for each dimensional feature vector. This learned model is used to assign the anonymous text to one of the known authors.

Keywords: Authorship attribution, Text classification, Support vector machine, Syntactic features, Lexical features, structural features, word ngrams and character ngrams.

1. Introduction:

Natural Language Processing is a research area that is used for many different purposes and it becomes more popular continuously. Authorship Recognition (AR) contains four major problems namely authorship attribution, authorship verification, authorship profiling and authorship clustering. AR is language dependant. So all the techniques developed for other languages need to be optimised for the Telugu language text. In this paper the point of interest is authorship attribution (AA) for Telugu text. AA can be defined in three ways. Firstly, for a given test document, find the author of the text from the defined set of authors. Secondly, for a given test document, believed to be written by one author from a set of authors then find which one, if any. Thirdly, for a given test

Peer-review under responsibility of scientific committee of International Conference on Computer, Communication and Convergence (ICCC 2015)
doi:10.1016/j.procs.2015.04.110

document, who is the author. There are two flavours of AA tasks: closed-class and open- class. The first definition is a closed class problem where as second and third definitions are open classes problems. In closed class problem the author to be identified is one from the given set of authors where as in open set problems the author to be identified may or may not in the defined author set. In this paper, it is addressed the author of an unknown text from the known author set which is a closed class problem.

Authors have their own style of speaking and writing. The writing style can be used as distinctive features to recognize its author. It can be considered as a typical classification problem, where a set of documents with known authors are used for training and the aim is to automatically determine the corresponding author of an anonymous text. In contrast to other classification tasks, it is not clear which features of a text should be used to classify an author.

In general, applications of AA include resolving historical questions of unclear or disputed authorship. In recent years, practical applications for author identification have grown in areas such as intelligence, criminal law, civil law, and computer security. AA has a long history with multiple application areas that include spam filtering [1], cyber bullying, plagiarism detection [2], author recognition of a given program [3], and web information management [2]. In forensic investigations where verifying the authorship of e-mails and newsgroup messages, or identifying the source of a piece of intelligence also considered as an AA applications.

The research on AA for Telugu language text has not attempted. Various features for extraction of author characteristics are attempted on different languages text, but not on Telugu text. Hence it is required to be thoroughly test the influence of different features of Telugu text for AA. In this paper an attempt is made on Telugu text for AA using different features and with their combination.

2. Related Work:

Authorship Analysis tasks can be broadly categorized into four ways. They are authorship attribution (AA), author profiling, author identification and clustering. AA can be termed as author identification as in [4, 5], closed-class task as in [6], categorization task as in [7], needle-in-a-haystack problem [8] and vanilla authorship attribution [7].

Text Categorization (TC) labels documents according to a set of predefined categories. Authorship Attribution (AA) can be viewed as a classification problem. The various steps in AA are pre-processing, feature extraction, feature selection and reduction, learning model generation from the selected features and finally measuring the performance of the learned model using various metrics.

Corresponding Author. Tel: 07893606134

E-Mail Address: nagkanna80@gmail.com

In [9] found that approximately 1,000 authorship features had been proposed in the literature. The various types of proposed features can be categorized into three types. They are syntactic, lexical and structural features. The various lexical features are content words, letter frequency, special characters , character n-grams, misspellings in [10,11], special use words in [12], 8 punctuation marks in [13], most frequent types [14], spelling errors, word form errors in [15], syntactically classified punctuation , syntactic structure [16], function word frequencies, POS trigrams or sequences of 3 in [17], word n-grams in [18], POS bigrams or sequences of 2 in [19], unigrams/types shared by training and testing samples in [20], 1024-character sequences in [21], content words, frequent words in [22], words bigrams or sequences in [23], emoticons, netabbrevs in [24], character n-grams or sequences in [25], non-function words in [26], frequency of lemmas (dictionary entry headwords), frequency of negative words in [27]. The various syntactic features are punctuation, function words [10], POS tags [11], syntactically classified punctuation, verbal phrases [15], syntactically classified punctuation [16], function word-token ratios [28], POS trigrams or sequences of 3 [17], punctuation frequency [18], POS bigrams or sequences of 2 [19], PCFG-obtained POS [26], syntactically classified punctuation [29], verbal phrases [30], phrase types, words per phrase type [31]. The structural features are font color, font size [10], word length distribution and vocabulary richness [11] , sentence length [12] , type-token ratio [28], phrase length [17], complexity measures applied to POS [19], function words [21] , hyperlinks, font formatting [29], punctuation distribution, word distribution [30]. In this paper it is addressed the problem of Authorship Attribution using various lexical, syntactic and structural features.

Section 3 describes the methodology adopted for author identification. The section 4 contains the experimental results and detailed discussion on the obtained results. The conclusions drawn from the discussions and possible feature extensions are mentioned in section 5.

3. Methodology:

The problem of Authorship Attribution is viewed as a Text Classification problem. In Text Classification, classes are identified based on the topic covered by the documents where as in Authorship Attribution classes are identified based on the documents written by the authors. For author identification the various steps need to be followed. They are firstly data pre-processing is required to perform for tokenize the input text and get the stemmed form of the tokens. In the second step, features are extracted from the known texts which can differentiate the authors writing style. As the extracted features may increase the dimensionality of the feature space, it is required to reduce the dimensionality space. As a third step feature selection measures are used to reduce the dimensionality of the feature space. After extracting the reduced feature vector space, in the fourth step these vectors are inputted to the classifier to obtain the learning model. In the fifth step, test document is inputted to the learning model to identify the author of unknown text.

The dataset contains 300 Telugu news articles written by 12 authors which were collected from the Telugu News articles. The training set contains 20 documents per author where as testing set contains 5 documents per author. The training set is used to create the learning model for each author. The learned model is used to identify the author for each text from the test set.

Till today the problem of Authorship Attribution is not yet attempted in the Indian context, especially on Telugu language text. There is a need to study the problem of Author identification as Indian languages are very rich in inflectional morphology [32]. Dravidian languages such as Telugu and Kannada are morphologically more complex compared with many languages in the world.

Data pre-processing is first and very important step in Authorship Attribution. Text documents in raw format are not suitable for pattern generation. So they need to be converted into a suitable input format. Data pre-processing involves tokenization, stop word removal and stemming. Tokenization is the process dividing the text into small units called tokens having useful semantic meaning. The unnecessary symbols like semicolons, colons, exclamation marks, hyphens, bullets, parenthesis, and numbers are removed from the raw text.

As in [33,34] a list of commonly repeated tokens appear in every text document such as pronouns, conjunctions and prepositions are removed as they do not have any effect on the classification process. This word list is identified using Telugu morphological Analyser (TMA) by tagging the Parts of speech (POS) for each word. Stemming is the process of removing prefixes and suffixes from tokens as in [35]. This process is used to reduce the number of variations of tokens. Telugu morphological analyser (TMA) is used to obtain the stemmed form for each token.

3.1 Feature Extraction:

In Authorship Attribution not only the text is important but also stylometry and other features that define the characteristics of a writer are more important. As a second step in Authorship Attribution various features are extracted from the pre-processed text. Features are grouped into three categories namely lexical, syntactic and structural features. The various lexical features considered in this experimental analysis are average number of words, average number of sentences, average syllables per word, average word length, and average sentence length as lexical features per author. These five style markers results to 5-dimetional feature vector.

The various vocabulary richness features considered are hapax legomena which is the number of words that only once occur in a given text, hapax dislegomena which is the number of words occurring twice, word parts of speech tag (POS) count and functional words i.e. stop words. The POS tagging was done using morphological analyzer. The words with POS tagging such as prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles are considered as stop words. The numbers of stop words in our corpus are 741 functional words. In total the numbers of 749 features are identified as vocabulary richness features. These features are used to form 749-dimentional feature vector space and it is termed as lvfv.

In this paper, word ngrams and character ngrams are considered as a separate set of features. It is considered that word bigrams feature vector as wbfv, word trigrams feature vector as wtfv, character bigram feature vector as cbfv and character trigram feature vector as ctfv. To reduce the dimensionality of the vector space for word ngram, it considered only the words with ngrams word count more than 45, similarly for character sequence with

ngram count more than 90 are considered. The dimensions of wbfv, wtfv, cbfv and ctfv individually are 125, 392, 549 and 1245 respectively. The dimensionality of combined word ngram count (wfv), character ngram count (cfv) and both word and character ngram count (wcfv) are 517, 1794 and 2311 respectively. When all these features are combined together it is obtained a vector with 3065-dimensions which is termed as tfv.

The third step in Authorship Attribution is feature selection. The extracted features from the previous step increases the dimensionality space of the input set. The machine learning classifiers suffers with the problem of curse of dimensionality as the dimensionality space increases. Hence it is required to remove irrelevant or not most relevant features from the feature set.

In this paper, chi-square metric as in [36] is used as a measure for feature selection, which is the most effective feature selection metric in the literature. Chi-square measures the correlation between feature and author set. Only features whose chi-square value is more than the threshold value were considered as features in the reduced feature vector. The relevance of feature t with the author set c is calculated as follows.

$$X^2(t,c) = \frac{N*(AD-BC)^2}{(A+C)*(B+D)*(A+B)*(C+D)}$$

where A is the number of times both feature t and author set c exists, B is the number of times feature t exists, but author set c doesn't exist, C is the number of times feature t doesn't exist, but author set c exists, D is the number of times both feature t and author set c doesn't exist, N be the total number of the training samples. As the value is more, the feature t is more relevant to the set c. Some of the features whose chi-square value is less than the threshold value are considered as non relevant to the class c. Using chi-square measure it is reduced the dimensionality for lvfv, wbfv, wtfv, cbfv, ctfv, wfv, cfv, wcfv and tfv as 86, 21, 32, 58, 104, 45,142, 161 and 264 and termed these reduced features as rlvfv, rwbfv, rwtfv, rcbfv, rctfv, rwfv, rcfv, rwcfv and rtfv respectively.

In the fourth step the learned model is generated for each feature vector using Support Vector Machine. Support vector machine (SVM) is proved to be an effective machine learning algorithms for text categorization. In [37] for AA, SVM is used to generate learning model by using lexical features such character n-grams and word n-grams to represent the text. SVM classifier is used to learn the boundaries between author sets where author sets are treated as classes. The learned model generated from the SVM is used for author identification of unknown text.

In the fifth step, author is assigned for a given test document. Test document is processed through the various steps. Test document feature vector is given as input to learned model. Then the learned model assigns one of the known authors to the test document.

4. Results and Discussions:

The performance of the various features and their combinations for Authorship Identification is measured using accuracy and F1 value. The accuracy and F1 measures are calculated as follows:

Accuracy is the number of text articles from test set for which the author is correctly assigned over the total number of articles in the test set as in Equation 1

$$Accuracy = \frac{Numberofdocumentsthatarecorrectlyassigned}{Totalnumberoftestdocuments} \quad (1)$$

F1 is calculated as in equation 2

$$F_1 = \frac{2 * precision * recall}{precision + recall} \quad (2)$$

Where

$$precision = \frac{Numberofdocumentscorrectlyauthorassigned}{Numberofdocumentsauthorassigned} \quad (3)$$

And

$$recall = \frac{Numberofdocumentscorrectlyassigned}{Totalnumberoftestdocuments} \quad (4)$$

The automatic author identification system's performance is measured using F1 metric and accuracy. The

F1 value is calculated using precision and recall as in equations 2,3 and 4. The accuracy is calculated as shown in equation 1. The best performance is obtained in terms of F1 metric and accuracy using support vector machine as a classifier are 89% and 91% respectively. The best performance is obtained when all the features such as lexical, vocabulary richness, word ngrams and character ngrams are combined together. The combination of character bigrams and character trigrams also results good performance both in terms of F1 and accuracy and the values are 85% and 87% respectively. The combination of character and word features with their bigrams and trigrams were also performed good. When considering only individual features, character trigram feature performance is good compared with all other features.

S.No.	Feature Vector	F1 value	Accuracy
1	Lexical-Vocabulary (lvfv)	0.68	0.74
2	Word Bigram (wbfv)	0.75	0.77
3	Word Trigram (wtfv)	0.67	0.71
4	Character Bigram (cbfv)	0.74	0.79
5	Character Trigram (ctfv)	0.81	0.84
6	Word Bi-Trigram (wfv)	0.72	0.75
7	Character Bi-Trigram (cfv)	0.85	0.87
8	Word-Character (wcfv)	0.82	0.85
9	All combined (tfv)	0.89	0.91

Table 1: The F1 and Accuracy for nine feature vectors before dimensionality reduction using SVM classifier

Features which were considered to learn the model for author identification may reduce the quality of the learned model. Similarly, the more number of features may slow down the process of authorship attribution. It is required to eliminate irrelevant features and not more relevant features from the original dimensionality space. In this paper, feature selection was carried out using chi-square measure on nine dimensional feature vectors. After the dimensionality reduction, the learned model is generated using support vector machine. The performance of the automatic authorship attribution model is calculated using F1 measure and accuracy. From the obtained results, it is observed that the learned model performance is good using a reduced dimensionality vector which combines all the features and the values are 82% for F1 measure and 87% for accuracy. The combination of character bigram and character trigram is also performing well with the accuracy as 81%. Character trigram was performed as a best individual feature vector among all the features with 79% of accuracy.

S.No.	Feature Vector	F1 value	Accuracy
1	Reduced Lexical-Vocabulary (lvfv)	0.62	0.65
2	Reduced Word Bigram (wbfv)	0.68	0.72
3	Reduced Word Trigram (wtfv)	0.63	0.67
4	Reduced Character Bigram (cbfv)	0.71	0.73
5	Reduced Character Trigram (ctfv)	0.75	0.79
6	Reduced Word Bi-Trigram (wfv)	0.65	0.70
7	Reduced Character Bi-Trigram (cfv)	0.79	0.81
8	Reduced Word-Character (wcfv)	0.76	0.80
9	Reduced All combined (tfv)	0.82	0.87

Table 2: The F1 and Accuracy for nine feature vectors after dimensionality reduction using SVM classifier

5. Conclusions:

In this paper, it was investigated the authorship attribution of Telugu text using various features such as average number of words, average number of sentences, average number of syllables per word, average word length, average sentence length, hapax legomena, hapax dislegomena, parts of speech tag(POS) count and functional words. It was also addressed the AA problem using bigrams, trigrams of word and character level features. It was also considered these features with their combinations. In these experiments these features are termed as lvfv, wbfv, wtfv, cbfv, ctfv, wfv, cfv, wcfv and tfv. The reduced dimensionality features of all these nine feature vectors are termed as rlvfv, rwbfv, rwtfv, rcbfv, rctfv, rwfv, rcfv, rwcfv and rtfv. A learned model is generated using support vector machine classifier for all these dimensional feature vectors. The performance of learned model for authorship identification is measured using F1 metric and accuracy. From the obtained results, it was concluded that the authorship attribution for Telugu text, character ngram features are more successful than all other features. The combination of word ngrams, character ngrams with lexical and vocabulary features were performed with good results than using the features separately. As part of feature work, this work can be extended for more number of features with various machine learning approaches. It can also be experimented with various feature selection approaches for feature vector dimensionality reduction.

6. References:

[1]De Vel, O., Anderson, A., Corney, M., Mohay, G.: Multi-topic e-mail authorship attribution forensics. In: Proceedings of the Workshop on Data Mining for Security Applications, 8th ACM Conference on Computer Security (2001)

[2] Stamatatos, E.: Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology (2011).

[3] Hayes, J.H.: Authorship attribution: A principal component and linear discriminant analysis of the consistent programmer hypothesis. I. J. Comput. Appl. pp. 79–99 (2008)

[4] Zheng, R., Li, J., Chen, H. & Huang, Z. (2006). A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. Journal of the American Society for Information Science and Technology, 57(3): 378-393.

[5] Chaski, C. E. (2007). The Keyboard Dilemma and Authorship Identification. In P. Craiger & S. Shenoi (Eds.), Advances in Digital Forensics III (pp. 133-146). Ney York, NY: Springer.

[6] Juola, P. (2008). Authoship Attribution. Hanover, MA: Now Publishers.

[7] Koppel , M., Schler, J., & Argamon, S. (2009). Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology, 60(1), 9-26.

[8] Koppel, M., Schler, J., & Messeri, E. (2008). Authorship Attribution in Law Enforcement Scenarios. In C.S. Gal, P. Kantor, & B. Saphira (Eds.), Security Informatics and Terrorism: Patrolling the Web (pp.111-119). Amsterdam: IOS.

[9] Rudman, J. (1998). The State of Authorship Attribution Studies: Some Problems and Solutions.
Computers and the Humanities, 31, 351-365.

[10] Abbasi, A., & Chen, H. (2005). Applying Authorship Analysis to Extremist-Group Web Forum
Messages. IEEE Intelligent Systems, 20(5), 67-75.

[11] Abbasi, A., & Chen, H. (2008). Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection. ACM Transactions on Information Systems, 26(2), 1-29.

[12] Argamon, S., Šarić, M., & Stein, S. S. (2003). Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[13] Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F. (2002). An Experiment in Authorship Attribution (pp. 29-37). Proceedings of JADT 2002: Sixth International Conference on Textual Data Statistical Analysis.

[14] Burrows, J. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing, 17(3), 267-86.

[15] Chaski, C. E. (2001). Empirical Evaluations of Language-Based Author Identification Techniques. Forensic Linguistics, 8(1), 1-65.

[16] Chaski, C. E. (2005). Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations.

International Journal of Digital Evidence, 4(1), 1-13.
[17] Gamon, M. (2004). Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features. Proceedings of the 20th International Conference on Computational Linguistics: Vol.4 (pp. 611-617). Stroudsburg, PA: Association for Computational Linguistics.
[18] Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. Literary and Linguistic Computing, 22(3), 425-442.
[19] Hirst, G., & Feiguina, O. (2007). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. Literary and Linguistic Computing, 22(4), 405-417.
[20] Jockers, M. L., Witten, D. M., & Criddle C. S. (2008). Reassessing Authorship of the Book of Mormon Using Delta and Nearest Shrunken Centroid Classification. Literary and Linguistic Computing, 23(4), 465-491.
[21] Juola, P., & Baayen, H. (2005). A Controlled –Corpus Experiment in Authorship Attribution by Cross-Entropy. Literary and Linguistic Computing, 20(1), 59-67.
[22] Koppel , M., Schler, J., & Argamon, S. (2009). Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology, 60(1), 9-26.
[23] Nazar, R., & Sánchez Pol, M. (2007). An Extremely Simple Authorship Attribution System. In M.T. Turell, J. Cicres, & M. S. Spassova (Eds.), Proceedings of the 2nd European IAFL Conference on Forensic Linguistics / Language and the Law 2006. Barcelona: Documenta Universitaria.
[24] Orebaugh, A., & Allnutt, J. (2009). Classification of Instant Messaging: Communications for Forensics Analysis. The International Journal of Forensic Computer Science, 4(1), 22- 28.
[25] Peng, F., Schuurmans, D., Keselj, V., & Wang, S. (2003). Language Independent Authorship Attribution Using Character Level Language Models. Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics: Vol. 1 (pp. 267-274). Stroudsburg, PA: Association for Computational Linguistics.
[26] Raghavan, S., Kovashka, A., & Mooney, R. (2010). Authorship Attribution Using Probabilistic Context-Free Grammars. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 38-42).
[27] Tambouratzis, G., & Vassiliou, M. (2007). Employing Thematic Variables for Enhancing Classification Accuracy Within Author Discrimination Experiments. Literary and Linguistic Computing, 22(2), 207-224.
[28] Corney, M. (2003). Analysing E-mail Text Authorship for Forensic Purposes. Unpublished MA thesis, Queensland University of Technology. Retrieved October 28, 2011
[29] Rico-Sulayes, A. (2011). Statistical Authorship Attribution of Mexican Drug Trafficking Online
Forum Posts. International Journal of Speech, Language and the Law, 18(1), 53-74.
[30] Spassova, M. S. (2008). Las Perífrasis Verbales del Español en la Atribución Forense de Autoría. In R. Monroy, and A. Sánchez (Eds.), 25 Años de Lingüística en España: Hitos y Retos. Actas del XXVI Congreso de AESLA (pp. 605-614). Murcia: Universidad de Murcia.
[31] Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-Based Authorship Attribution without Lexical Measures. Computers and the Humanities, 35, 193-214.
[32] B.Vishnu Vardhan, L. Pratap Reddy, A. Vinay Babu, " A Model for Overlapping Trigram technique for Telugu Script" Journal of Theoritical and Applied Information Technology, Vol. 3, No. 3, Sep. 2007, pp 9-14.
[33] B.Vishnu Vardhan,P.Vijaypal Reddy,A.Govardhan"Analysis of BMW model for title word selection on Indic scripts", International Journal of Computer Application (IJCA) Vol 18 Number 8 March 2011 pp 21-25
[34] B.Vishnu Vardhan,P.Vijaypal Reddy, A.Govardhan"Corpus based Extractive summarization for Indic script", International Conference on Asian Language Processing (IALP) IEEE Computer Society (IALP 2011) pp 154-157
[35] P. Vijay pal Reddy, Vishnu Murthy.G, Dr. B. Vishnu Vardhan, K. Sarangam "A comparative study on term weighting methods for automated telugu text categorization with effective classifiers" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.6, November 2013
[36] D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. The American Physical Society, 88(4):048702, 2002.
[37] Stamatatos, E.: Author Identification: Using Text Sampling to Handle the Class Imbalance Problem. Information Processing and Management, 44(2), pp. 790--799 (2008).